第九讲

王伟文　暨南大学

1　度量空间

度量空间 $(\mathcal{T},\rho)$ 由非空集合 $\mathcal{T}$ 及度量 $\rho:\mathcal{T}\times\mathcal{T}\to\mathbb{R}$ 组成, 其中度量 ρ 满足

- 非负性: $\forall(\theta,\tilde{\theta})\in\mathcal{T}\times\mathcal{T},\rho(\theta,\tilde{\theta})\geq 0$. $\rho(\theta,\tilde{\theta})=0$ 当且仅当 $\theta=\tilde{\theta}$.
- 对称性: $\forall(\theta,\tilde{\theta})\in\mathcal{T}\times\mathcal{T},\rho(\theta,\tilde{\theta})=\rho(\tilde{\theta},\theta)$.
- 三角不等式: $\forall(\theta,\tilde{\theta},\hat{\theta})\in\mathcal{T}^3$

$$\rho(\theta,\tilde{\theta})\leq\rho(\theta,\hat{\theta})+\rho(\hat{\theta},\tilde{\theta}).$$

例 1.1. $(\mathbb{R}^d,\rho)$, 欧氏距离 $\rho(\theta,\tilde{\theta})=\sqrt{\sum_{i=1}^{d}(\theta_i-\tilde{\theta}_i)^2}$.

例 1.2. $(\{0,1\}^d,\rho)$, 汉明 (Hamming) 距离 $\rho(\theta,\tilde{\theta})=\frac{1}{d}\sum_{i=1}^{d}\mathbb{I}_{\{\theta_i=\tilde{\theta}_i\}}$

例 1.3. $(C[0,1],\rho)$, 其中 $C[0,1]$ 表示所有在 $[0,1]$ 上连续的函数的集合

$$\rho(f,g)=\sup_{x\in[0,1]}|f(x)-g(x)|$$

接下来我们将回答如何“测量”一个度量空间的“大小”.

定义 1.1: 覆盖数 $\mathcal{N}(\delta;\mathcal{T},\rho)$

给定集合 $\mathcal{C}=\{\theta^1,\theta^2,\ldots,\theta^n\}\subset\mathcal{T}$ 及常数 $\delta>0$. 若 $\forall\theta\in\mathcal{T}$, $\exists i\in[n]$ 使得

$$\rho(\theta,\theta^i)\leq\delta,$$

则称 $\mathcal{C}$ 为集合 $\mathcal{T}$ 在度量 ρ 下的 δ-覆盖. 集合 $\mathcal{T}$ 的 δ-覆盖数 $\mathcal{N}(\delta;\mathcal{T},\rho)$ 即为 $\mathcal{T}$ 的最小 δ-覆盖的元素个数, 其中“最小”指元素个数最少.

评论. 若 $\mathcal{C}=\{\theta^1,\theta^2,\ldots,\theta^n\}$ 为 $\mathcal{T}$ 的 δ-覆盖, 则

$$\mathcal{T}\subseteq\bigcup_{i\in[n]}\mathcal{B}_\rho(\theta^i;\delta)$$

其中 $\mathcal{B}_\rho(\theta^i;\delta)$ 为在度量 ρ 下以 δ 为半径, θ^i 为中心的球.

例 1.4. 考虑闭区间 $\mathcal{T}=[-1,1]$, 定义度量 $\rho(\theta,\tilde{\theta})=|\theta-\tilde{\theta}|$, 则

$$\mathcal{N}(\delta;\mathcal{T},\rho)\leq 2+\frac{1}{\delta}.$$

定义集合 $\mathcal{C}=\{-1,-1+2\delta,-1+4\delta,\cdots,-1+2n\delta,1\}\subset\mathcal{T}$, 其中 $n=\lfloor\frac{1}{\delta}\rfloor$, 容易知道 $\mathcal{C}$ 为 $[-1,1]$ 上的 δ-覆盖.

$$\mathtt{card}(C)=\lfloor\frac{1}{\delta}\rfloor+2\leq\frac{1}{\delta}+2,$$

因此

$$\mathcal{N}(\delta;\mathcal{T},\rho)\leq 2+\frac{1}{\delta}.$$

推广至 d-维空间,

$$\mathcal{N}(\delta;[-1,1]^d,\|\cdot\|_\infty)\leq\left(2+\frac{1}{\delta}\right)^d.$$

定义 1.2: 填充数 (Packing number)$\mathcal{M}(\delta;\mathcal{T},\rho)$

给定集合 $\mathcal{P}=\{\theta^1,\theta^2,\ldots,\theta^m\}\subset\mathcal{T}$ 及常数 $\delta>0$, 若对任意 $i,j\in[m]$ 且 $i\neq j$ 有

$$\rho(\theta^i,\theta^j)>\delta$$

则称 $\mathcal{P}$ 为度量 ρ 下集合 $\mathcal{T}$ 的 δ-填充 (packing). 集合 $\mathcal{T}$ 的填充数 $\mathcal{M}(\delta;\mathcal{T},\rho)$ 为最大 δ-填充的元素个数, 其中“最大”指元素个数最多.

引理 1.1

给定度量空间 $(\mathcal{T},\rho)$, 对任意 $\delta>0$,

$$\mathcal{M}(2\delta;\mathcal{T},\rho)\overset{(a)}{\leq}\mathcal{N}(\delta;\mathcal{T},\rho)\overset{(b)}{\leq}\mathcal{M}(\delta;\mathcal{T},\rho)$$

证明. ***b.*** 这里只需证明 $\mathcal{T}$ 的最大 δ-填充同时也是 δ-覆盖.

记 $\mathcal{P}=\{\theta^1,\theta^2,\ldots,\theta^m\}\subset\mathcal{T}$ 为 $\mathcal{T}$ 的一个最大 δ-填充, 其中 $m=\mathcal{M}(\delta;\mathcal{T},\rho)$.

若 $\mathcal{P}$ 不是集合 $\mathcal{T}$ 的 δ-覆盖, 则存在 $\tilde{\theta}\in\mathcal{T}$ 使得

$$\rho(\tilde{\theta},\theta^i)>\delta\quad\forall i\in[m].$$

故 $\mathcal{P}\bigcup\{\tilde{\theta}\}$ 是 $\mathcal{T}$ 的一个 δ-填充, 与 $\mathcal{P}$ 为最大 δ-填充矛盾.

a. 这里需要证明任意 2δ-填充的元素个数总是不超过任意 δ-覆盖的元素个数.

记 $\mathcal{P}=\{\theta^1,\theta^2,\ldots,\theta^m\}\subset\mathcal{T}$ 为 $\mathcal{T}$ 的 2δ-填充, $\mathcal{C}=\{\tilde{\theta}^1,\tilde{\theta}^2,\ldots,\tilde{\theta}^n\}\subset\mathcal{T}$ 为 $\mathcal{T}$ 的 δ-覆盖.

因为 $\mathcal{P}\subset\mathcal{T}$, $\mathcal{T}$ 为 δ-覆盖, 故 $\forall\theta^i\in\mathcal{P}$, $\exists\tilde{\theta}^j\in\mathcal{C}$ 使得

$$\rho(\theta^i,\tilde{\theta}^j)\leq\delta.$$

同时, 若 $\forall\theta^k \in \mathcal{P}$ 且 $\theta^k \neq \theta^i$, 有 $\rho(\theta^k, \tilde{\theta}^j) > \delta$, 否则

$$\rho(\theta^i, \theta^k) \leq \rho(\theta^i, \tilde{\theta}^j) + \rho(\theta^k, \tilde{\theta}^j) \leq 2\delta$$

与 $\mathcal{P}$ 为 2δ-填充矛盾.

上述关系表明, 对于任意 $\theta^i \in \mathcal{P}$, 存在唯一 $\tilde{\theta}^j \in \mathcal{C}$ 满足

$$\rho(\theta^i, \tilde{\theta}^j) \leq \delta$$

由鸽巢原理知

$$\texttt{card}(\mathcal{P}) \leq \texttt{card}(\mathcal{C})$$

因此

$$\mathscr{M}(2\delta; \mathcal{T}, \rho) \leq \mathscr{N}(\delta; \mathcal{T}, \rho)$$

□

例 1.5. 回顾例 1. 4. 考虑闭区间 $\mathcal{T} = [-1, 1]$, 度量 $\rho(\theta, \tilde{\theta}) = |\theta - \tilde{\theta}|$. 定义集合 $\mathcal{P} = \{-1, -1 + 2(1+\epsilon)\delta, -1 + 4(1+\epsilon)\delta, \cdots, -1 + 2(n-1)(1+\epsilon)\delta\} \subset \mathcal{T}$, 其中 $n = \lfloor \frac{1}{\delta} \rfloor$, $\delta < 1$, $0 < \epsilon < \frac{\delta}{1-\delta}$, 容易知道 $\mathcal{P}$ 为 $[-1, 1]$ 上的 2δ-填充.

$$-1 + 2(n-1)(1+\epsilon)\delta < -1 + 2(\frac{1}{\delta} - 1)(1 + \frac{\delta}{1-\delta})\delta = 1$$
$$\rho(\theta^i, \theta^j) \geq 2(1+\epsilon)\delta > 2\delta \quad \forall \theta^i, \theta^j \in \mathcal{P}.$$

由引理 1.1 知

$$\lfloor \frac{1}{\delta} \rfloor \leq \mathscr{M}(2\delta; \mathcal{T}, \rho) \leq \mathscr{N}(\delta; \mathcal{T}, \rho) \leq \lfloor \frac{1}{\delta} \rfloor + 2.$$

若 δ 充分小

$$\log \mathscr{N}(\delta; \mathcal{T}, \rho) \asymp \log(\frac{1}{\delta})$$

推广至 d 维空间

$$\log \mathscr{N}(\delta; [-1, 1]^d, \| \cdot \|_\infty) \asymp d \log(\frac{1}{\delta})$$

集合的 Minkowski 和

$$\mathcal{A} + \mathcal{B} = \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in \mathcal{A}, \boldsymbol{y} \in \mathcal{B}\}$$

定理 1.1

考虑 $\mathbb{R}^d$ 上的两个范数 $\|\cdot\|$ 及 $\|\cdot\|'$, 定义 $\mathcal{B}=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}\|\le 1\}$ 及 $\mathcal{B}'=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}\|'\le 1\}$, 在度量 $\|\cdot\|'$ 下 $\mathcal{B}$ 的 δ-覆盖数满足

$$\left(\frac{1}{\delta}\right)^d\frac{\texttt{vol}(\mathcal{B})}{\texttt{vol}(\mathcal{B}')}\overset{(a)}{\le}\mathscr{N}(\delta;\mathcal{B},\|\cdot\|')\overset{(b)}{\le}\frac{\texttt{vol}(\frac{2}{\delta}\mathcal{B}+\mathcal{B}')}{\texttt{vol}(\mathcal{B}')}$$

证明. $\boldsymbol{a}$. 设 $\mathcal{C}=\{\theta^1,\cdots,\theta^n\}\subseteq\mathcal{B}$ 为 $\mathcal{B}$ 在 $\|\cdot\|'$ 下的 δ-覆盖, 故 $\forall\theta\in\mathcal{B}$, $\exists\theta^i\in\mathcal{C}$, 使得

$$\|\theta-\theta^i\|'\le\delta,$$

从而 $\theta-\theta^i\in\delta\mathcal{B}'$, $\theta\in\theta^i+\delta\mathcal{B}'$, 因此

$$\mathcal{B}\subseteq\bigcup_{i\in[n]}\{\theta^i+\delta\mathcal{B}'\}$$

故有

$$\texttt{vol}(\mathcal{B})\le\sum_{i=1}^n\texttt{vol}(\theta^i+\delta\mathcal{B}')=n\texttt{vol}(\delta\mathcal{B}')=n\delta^d\texttt{vol}(\mathcal{B}')$$

即

$$\left(\frac{1}{\delta}\right)^d\frac{\texttt{vol}(\mathcal{B})}{\texttt{vol}(\mathcal{B}')}\le n$$

上式对任意 δ-覆盖均成立, 故

$$\left(\frac{1}{\delta}\right)^d\frac{\texttt{vol}(\mathcal{B})}{\texttt{vol}(\mathcal{B}')}\le\mathscr{N}(\delta;\mathcal{B},\|\cdot\|')$$

$\boldsymbol{b}$. 设 $\mathcal{P}=\{\tilde{\theta}^1,\tilde{\theta}^2,\ldots,\tilde{\theta}^m\}$ 为 $\mathcal{B}$ 在 $\|\cdot\|'$ 下的最大 δ-填充, , 定义 $\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)$, $i\in[m]$, 因为

$$\|\tilde{\theta}^i-\tilde{\theta}^j\|>\delta,$$

所以

$$\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)\bigcap\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^j,\delta)\quad\forall\tilde{\theta}^i,\tilde{\theta}^j\in\mathcal{P}.$$

接下来证明 $\bigcup_{i\in[m]}\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)\subseteq\mathcal{B}+\frac{\delta}{2}\mathcal{B}'$.

$\forall\theta\in\bigcup_{i\in[m]}\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)$, $\exists\tilde{\theta}^i\in\mathcal{P}$ 使得

$$\theta\in\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)\Longrightarrow\|\theta-\tilde{\theta}^i\|'\le\frac{\delta}{2}$$

同 ($\boldsymbol{a}$) 的证明, $\theta\in\tilde{\theta}^i+\frac{\delta}{2}\mathcal{B}'$, 又因为 $\tilde{\theta}^i\in\mathcal{B}$, 故 $\theta\in\mathcal{B}+\frac{\delta}{2}\mathcal{B}'$.

因此

$$\texttt{vol}(\bigcup_{i\in[m]}\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)) = \sum_{i=1}^m \texttt{vol}(\mathcal{B}_{\|\cdot\|'}(\tilde{\theta}^i,\delta)))$$
$$= m\texttt{vol}(\frac{\delta}{2}\mathcal{B}')$$
$$\le \texttt{vol}(\mathcal{B}+\frac{\delta}{2}\mathcal{B}')$$

整理得到

$$m \le \frac{\texttt{vol}(\mathcal{B}+\frac{\delta}{2}\mathcal{B}')}{\texttt{vol}(\frac{\delta}{2}\mathcal{B}')} = \frac{\texttt{vol}\left(\frac{2}{\delta}\mathcal{B}+\mathcal{B}'\right)}{\texttt{vol}(\mathcal{B}')}$$

由引理 1.1

$$\mathscr{N}(\delta;\mathcal{B},\|\cdot\|') \le m \le \frac{\texttt{vol}\left(\frac{2}{\delta}\mathcal{B}+\mathcal{B}'\right)}{\texttt{vol}(\mathcal{B}')}.$$

□

例 1.6. 取 $\|\cdot\| = \|\cdot\|' = \|\cdot\|_\infty$, 此时 $\mathcal{B}=\mathcal{B}'$, 对于度量空间 $(\mathcal{B}_{\|\cdot\|_\infty}(\mathbf{0},1),\|\cdot\|_\infty)$, 注意到 $\mathcal{B}_{\|\cdot\|_\infty}(\mathbf{0},1)=[-1,1]^d$, 应用定理 1.1 可得到

$$\left(\frac{1}{\delta}\right)^d \le \mathscr{N}(\delta;[-1,1]^d,\|\cdot\|_\infty) \le \left(\frac{2}{\delta}+1\right)^d$$
$$d\log\frac{1}{\delta} \le \log\mathscr{N}(\delta;[-1,1]^d,\|\cdot\|_\infty) \le d\log(\frac{2}{\delta}+1)$$

2 Gaussian 复杂度和 Rademacher 复杂度

给定集合 $\mathcal{T}\subseteq\mathbb{R}^d$, 定义典则 Gaussian 过程

$$G_\theta = \langle\theta,\boldsymbol{w}\rangle \quad \forall\theta\in\mathcal{T}, w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

称

$$\mathscr{G}(\mathcal{T}) = \mathbb{E}_{\boldsymbol{w}}[\sup_{\theta\in\mathcal{T}}\langle\theta,\boldsymbol{w}\rangle]$$

为集合 $\mathcal{T}$ 的 Gaussian 复杂度.

类似地, 定义 Rademacher 过程

$$R_\theta = \langle\theta,\boldsymbol{\epsilon}\rangle \quad \forall\theta\in\mathcal{T},$$

其中 $\boldsymbol{\epsilon}=(\epsilon_i)_{i\in[d]}$ 各元素为相互独立的 Rademacher 变量.

称

$$\mathscr{R}(\mathcal{T}) = \mathbb{E}_{\boldsymbol{\epsilon}}[\sup_{\theta\in\mathcal{T}}\langle\theta,\boldsymbol{\epsilon}\rangle]$$

为集合 $\mathcal{T}$ 的 Rademacher 复杂度.

例 2.1. 考虑集合 $\mathcal{B}_2^d(1)=\{\theta\in\mathbb{R}^d:\|\theta\|_2\leq 1\}$, 计算 $\mathscr{R}(\mathcal{B}_2(1))$ 和 $\mathscr{G}(\mathcal{B}_2^d(1))$.

$$
\begin{aligned}
\mathscr{R}(\mathcal{B}_2^d(1)) &= \mathbb{E}_{\boldsymbol{\epsilon}}[\sup_{\theta\in\mathcal{B}_2^d(1)}\langle\theta,\boldsymbol{\epsilon}\rangle] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}[\left\langle\frac{\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|_2},\boldsymbol{\epsilon}\right\rangle] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{\epsilon}\|_2 \\
&= \mathbb{E}\sqrt{\sum_{i=1}^d\epsilon_i^2}=\sqrt{d}
\end{aligned}
$$

类似地,

$$
\mathscr{G}(\mathcal{B}_2^d(1))=\mathbb{E}\|\boldsymbol{w}\|_2=\mathbb{E}\sqrt{\sum_{i=1}^d w_i^2}\leq\sqrt{\sum_{i=1}^d\mathbb{E}w_i^2}=\sqrt{d}
$$

例 2.2. 考虑一致 b-有界函数类 $\mathcal{F}$, 即 $\forall f\in\mathcal{F}$, $\|f\|_\infty\leq b$. 给定 n 个数据点 $x_1^n=(x_i)_{\{i\in[n]\}}$, 定义集合

$$
\mathcal{F}(x_1^n)/n=\{(f(x_1),f(x_2),\cdots,f(x_n))/n:f\in\mathcal{F}\},
$$

计算 $\mathscr{G}(\mathcal{F}(x_1^n)/n)$.

$$
\begin{aligned}
\mathscr{G}(\mathcal{F}(x_1^n)/n) &= \mathbb{E}_{\boldsymbol{w}}[\sup_{f\in\mathcal{F}}\langle f(x_1^n)/n,\boldsymbol{w}\rangle] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\frac{f(x_i)w_i}{n}\right] \\
&\leq \mathbb{E}_{\boldsymbol{w}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sqrt{\sum_i^n(f(x_i))^2}\|\boldsymbol{w}\|_2\right] \\
&\leq \mathbb{E}_{\boldsymbol{w}}\frac{b}{\sqrt{n}}\|\boldsymbol{w}\|_2\leq\frac{b}{\sqrt{n}}\sqrt{n}=b.
\end{aligned}
$$

引理 2.1

对任意 $\mathcal{T}\subseteq\mathbb{R}^d$

(1) $\mathscr{R}(\mathcal{T})\leq\sqrt{\frac{\pi}{2}}\mathscr{G}(\mathcal{T})$,

(2) $\mathscr{G}(\mathcal{T})\leq 2\sqrt{\log d}\mathscr{R}(\mathcal{T})$.